Mall Customer Segmentation

### Vedant Bohat

[*vedantbohat@gmail.com*](mailto:vedantbohat@gmail.com)

# Abstract

Many consumers buy products from the mall, and in order for the mall to generate more revenue, the authorities must attract these customers, which requires a big amount of capital. The yield is just around 30-40% after the advertisement. As a result, customer segmentation enters the picture.

Customer segmentation is a popular unsupervised learning application, and we'll use it to focus just on potential clients (customers whose probability of buying the product is very high). The output will be dramatically increased to 90-95 percent using this strategy.

This project aims to build clusters of customers based on their Spending Score and Annual Income. The algorithm used in this project is K-means.

# Table of Contents

|  |  |
| --- | --- |
| Topic | Page No. |
| 1. Introduction.………………………………………...…………...…………...……… | 2 |
| 2. Dataset………………………………………….……………………………….…. | 2 |
| 2.1 Snapshot of Dataset...……………………………....………………………. ….. | 2 |
| 3. Proposed Method and Architecture. …..………….….………………………….. | 3 |
| 3.1 Data Science Project Architecture……………………………………….…….... | 3 |
| 4. Methodology.………………………………….……………………………………. | 4 |
|  |  |
| 5. Implementation and Data Analysis.……………………………………………  5.1 Gender Plot….…………………...…………………………………………. | 4  4 |
| 5.2 Age Plot.…………………………………………………………………...... | 5 |
| 5.3 Age Vs Spending Score..……………………...……………………………. | 5 |
| 5.4 Annual Income Vs Spending Score……………….………………………… | 5 |
| *6.* Conclusion...…………………………………………………………………………. | 6 |
| 6.1 Annual Income Vs Spending Score after Clustering………………….……….. |  |

1. **Introduction**

One of the most common applications of unsupervised learning is to create predictions and locate clusters of potential mall clients in order to find relevant strategies to boost the mall's revenue.

For example, if a group of customers has a high income but a low spending score (amount spent in the mall), we can convert them into future customers (with a high spending score) by employing methods such as better advertising, accepting feedback, and enhancing product quality.

To identify such customers, this project analyses and forms clusters based on different criteria which are discussed in the further sections.

# Dataset

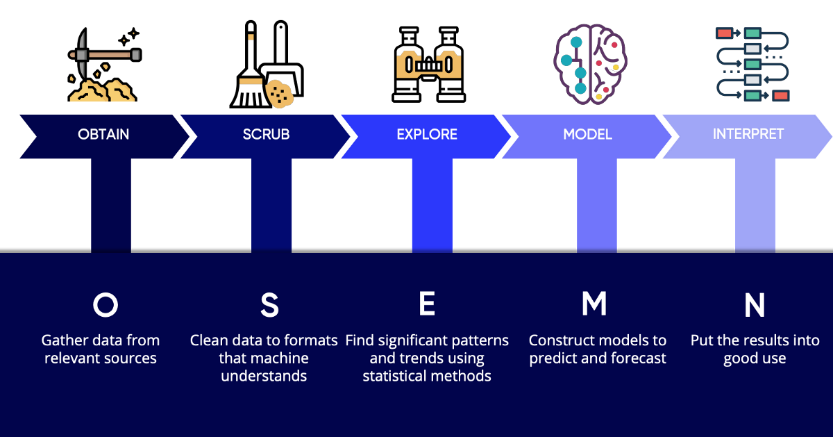
The dataset name is ‘Mall\_Customers.csv’ consists of 5 columns which are CustomerID, Gender, Age, Annual Income (k$), Spending Score (1-100) where Gender is a categorical value and rest all features are numeric.



*2.1 Snapshot of Dataset*

The size of the dataset is (200, 5) which is 200 rows and 5 columns.

# Proposed Method and Architecture

**

*3.1 Data Science Project Architecture*

## Problem Statement

Customer Segmentation is a popular application of unsupervised learning. Using clustering, identify segments of customers to target the potential user base. They divide customers into groups according to common characteristics like gender, age, interests, and spending habits so they can market to each group effectively.

Use K-means clustering and also visualize the gender and age distributions. Then analyze their annual incomes and spending scores.

## Data

The size of the dataset is (200, 5) which is 200 rows and 5 columns. Also, the dataset does not contain any NULL or NaN values.

## Algorithms

K-means algorithm is used in this project to analyze and form clusters of customers based on their income and spending score features.

## Model

K-means model is used and is hyper tuned parameters like *n\_clusters=5* using elbow method to find the optimal number of clusters also *init=’k-means++’* to avoid random initialization traps.

## Programming and Environment

Programming Language: Python 3.10

Environment (Libraries and Technologies): NumPy, Pandas, Matplotlib, Seaborn, Jupyter Notebook.

# Methodology

The Data Science Methodology aims to answer basic questions in a prescribed sequence, that cover the five main aspects of data science projects. These aspects are:

* From Problem to Approach
* From Requirements to Collection
* From Understanding to Preparation
* From Modelling to Evaluation
* From Deployment to Feedback

In this project, the prescribed sequence is:

* Creating an approach to solve the given problem statement
* Exploring the dataset and obtaining useful insight from the same
* Cleaning the dataset by handling nan values, remove duplicate records, etc.
* Data Visualization used to obtain important information from the data
* Data Preprocessing is performed to make the data ready to fit the model this includes feature scaling, splitting the dataset into features and labels, etc.
* Model Building

# Implementation and Data Analysis

On performing data visualization on the dataset, a lot of insights were obtained such as:

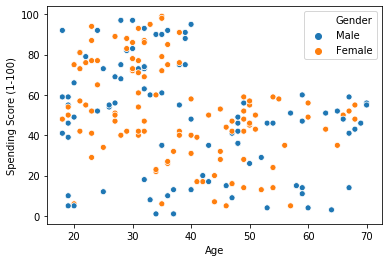
### Gender Plot Analysis

From the Count plot, it is observed that the number of Female customers is more than the total number of Male customers.

*5.1 Gender Plot*

### Age Plot Analysis

From the Histogram it is evident that there are 3 age groups that are more frequently shop at the mall, they are: 15-22 years, 30-40 years, and 45-50 years.

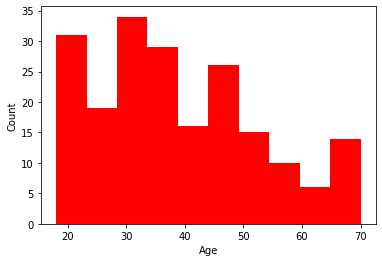


*5.3 Age Vs Spending Score*

### Annual Income Vs Spending Score

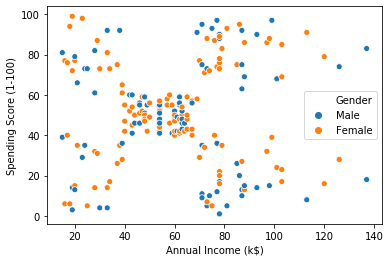
**Analysis**

We observe that there are 5 clusters and can be categorized as:



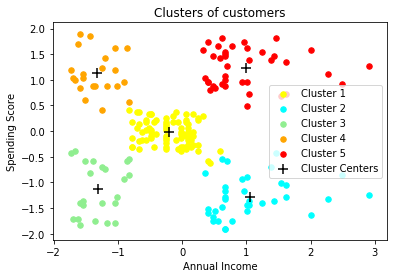
*5.2 Age Plot*

### Age Vs Spending Score Analysis

1. From the Age Vs Spending Score plot we observe that customers whose spending score is more than 65 have their Age in the range of 15-42 years. Also from the Scatter plot it is observed that customers whose spending score is more than 65 consists of more Females than Males.
2. The customers having average spending score ie: in the range of 40-60 consists of the age group of the range 15-75 years and the count of males and females in this age group is also approximately the same.
   1. High Income, High Spending Score (Top Right Cluster)
   2. High Income, Low Spending Score (Bottom Right Cluster)
   3. Average Income, Average Spending Score (Center Cluster)
   4. Low Income, High Spending Score (Top Left Cluster)
   5. Low Income, Low Spending Score (Bottom Left Cluster) *5.4 Annual Income Vs Spending Score*

# Conclusion

For this project, the K-means algorithm is used and performs the best (with n\_clusters = 5 and init = ‘kmeans++’). After the clustering algorithm is applied to the dataset, this is the output.



*6.1 Annual Income Vs Spending Score after Clustering*

### Clustering Analysis

1. High Income, High Spending Score (Cluster 5) - Target these customers by sending new product alerts which would lead to an increase in the revenue collected by the mall as they are loyal customers.
2. High Income, Low Spending Score (Cluster 2) - Target these customers by asking the feedback and advertising the product in a better way to convert them into Cluster 5 customers.
3. Average Income, Average Spending Score (Cluster 1) - Can target this set of customers by providing them with Low-cost EMI's, etc.
4. Low Income, High Spending Score (Cluster 4) - May or may not target these groups of customers based on the policy of the mall.
5. Low Income, Low Spending Score (Cluster 3) - Don't target these customers since they have less income and need to save money.